

רגרסיה ליניארית פשוטה

רגרסיה ליניארית פשוטה מסתמכת על המתאם הליניארי בין המשתנה התלוי (המנובא) לב"ת (המנבא).

מקדם המתאם:

$$r = \frac{\text{cov}(x, y)}{S_x \cdot S_y} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)} \cdot \sqrt{\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}} = \frac{S_{XY}}{\sqrt{S_{XX}} \cdot \sqrt{S_{YY}}}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
 המודל באוכלוסיה:

כאשר:

β_0 הוא החותך

β_1 הוא שיפוע

ε_i הינו גורם הטעות מסביב לקו הליניארי.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
 המודל הנאמד (על סמך מדגם):

נמחיש בדוגמא:

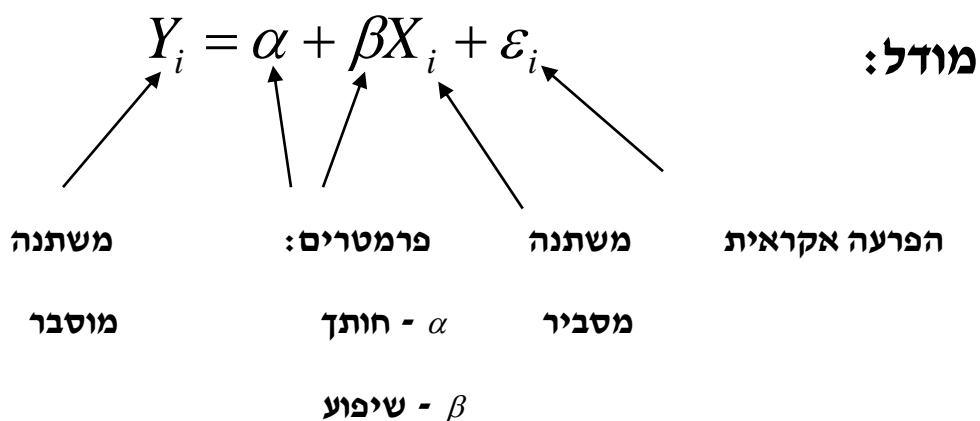
מתווך דירות בתל אביב רצה לבדוק איך משפיע גודלה של דירה על המחיר שבו היא נמכרת.

הוא הניח 2 הנחות מקדימות:

1) רק גודל הדירה משפיע על מחיר הדירה באופן שיטתי. כל שאר הדברים המשפיעים על מחיר הדירה הם אקראיים ולא ניתנים לחיזוי.

2) ההשפעה של גודל הדירה על מחיר הדירה היא ליניארית.

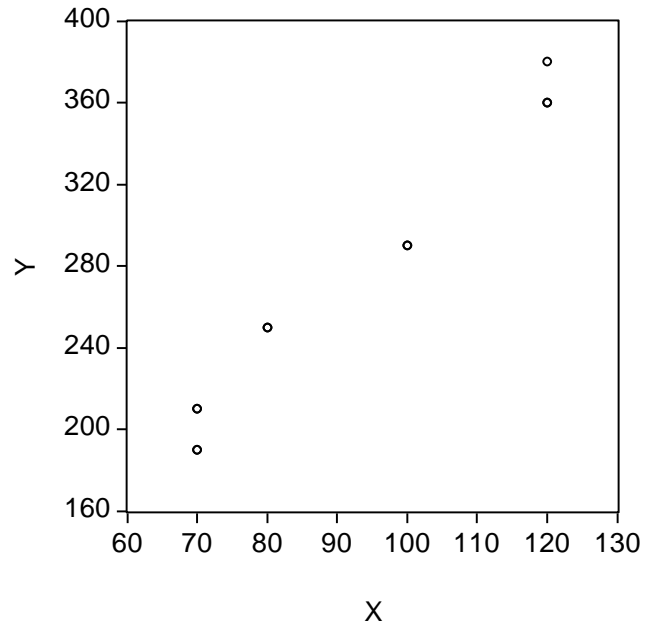
שתי ההנחות האלה מאפיינות את הקשר. אם נסמן את גודל הדירה ב- X ואת מחיר הדירה ב- Y , נוכל לכתוב באופן מתמטי כי $y_i = \alpha + \beta x_i + \varepsilon$. זהו המודל של המתווך. X ו- Y הם המשתנים של המודל. Y הוא המשתנה המוסבר של המודל. X הוא המשתנה המסביר של המודל (יכול להיות יותר ממשתנה מסביר אחד). α ו- β הם הפרמטרים של המודל. α נקרא חותך. β , או כל מקדם אחר של משתנה מסביר, נקרא שיפוע. ε מכונה הפרעה האקראית.



אחרי הגדרת המודל המתווך אסף נתונים על 6 דירות, שנמכרו בחודש האחרון באותו איזור. זהו המדגם של המתווך. במדגם יש 6 תצפיות. נוהגים להציג את המודל כאשר לכל משתנה נוסף אינדקס $y_i = \alpha + \beta x_i + \varepsilon_i$. האינדקס מייצג את מספר התצפית.

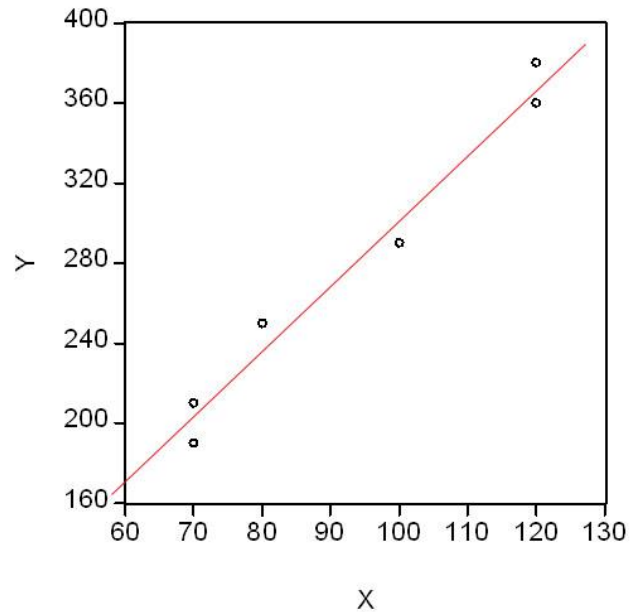
מספר הדירה	גודל הדירה במ"ר	מחיר הדירה באלפי דולרים
1	$X_1 = 70$	$Y_1 = 190$
2	$X_2 = 70$	$Y_2 = 210$
3	$X_3 = 80$	$Y_3 = 250$
4	$X_4 = 100$	$Y_4 = 290$
5	$X_5 = 120$	$Y_5 = 360$
6	$X_6 = 120$	$Y_6 = 380$

נציג את 6 התצפיות בגרף:



מהו הקו הישר המתאר את הקשר בין שני המשתנים בצורה הטובה ביותר? (הקו הוא ישר בגלל שהמתווך הניח לינאריות של המודל).

מסתבר שקו הרגרסיה הטוב ביותר הוא קו שחושב בשיטת הריבועים הפחותים (הסבר בהמשך):



הנוסחה של הקו היא: $\hat{Y}_i = -27.32 + 3.29 X_i$.

זהו כנראה לא הקו האמיתי, אך ממילא את הקו האמיתי אף פעם אי אפשר לדעת. סביר שקו זה הוא די קרוב לקו האמיתי.

לפי הנוסחה כל מ"ר נוסף שיש בדירה מעלה את מחירה ב-3,290 דולר.

מקו זה יודע המתווך להעריך מחירים של דירות. כשפנה אליו בעל דירה שגודלה 90 מ"ר ושאל אותו מה שווי הדירה, חישב המתווך לפי הנוסחה, $-27.32 + 3.29 \cdot 90 = 268.78$, והשיב לבעל הדירה: "המחיר שאתה יכול לקבל עליה הוא 268,780 דולר. אם יהיה לך מזל תקבל יותר, אבל יכול להיות שתצטרך למכור בפחות".

כלומר נוכל לומר כי אם יהיה לו מזל אז ההפרעה האקראית תהיה חיובית, ואם לא – היא תהיה שלילית.

לסיכום:

1) במודל $y_i = \alpha + \beta x_i + \varepsilon_i$, α ו- β הם מספרים קבועים אך לא ידועים. אנו יכולים להעריך אותם ולקבל אומדים (תהליך קבלת האומדנים נקרא אמידה).

2) $\hat{\alpha}$ הוא האומד ל- α . $\hat{\beta}$ הוא האומד ל- β .

3) אומדי ריבועים פחותים (אר"פ) הם אומדים שחושבו בשיטת הריבועים הפחותים. אומדי הריבועים הפחותים מסומנים בד"כ ע"י 'קובע' - $\hat{\beta}, \hat{\alpha}$.

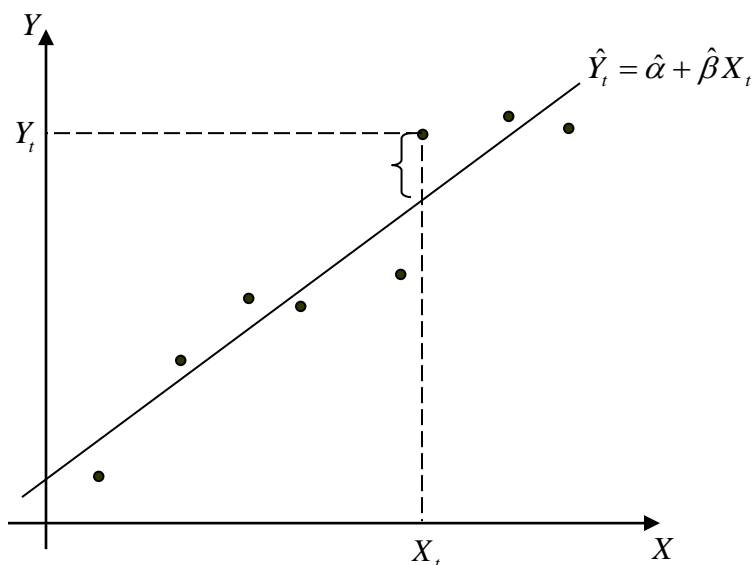
4) בעוד α ו- β הם קבועים, $\hat{\alpha}$ ו- $\hat{\beta}$ הם משתנים מקריים. מדוע? מפני שבכל מדגם מתקבלים $\hat{\alpha}$ ו- $\hat{\beta}$ אחרים.

5) את α ו- β אי אפשר לדעת, ולכן אי אפשר לדעת מהו הקו האמיתי, וכן אי אפשר לדעת את ε .

6) אפשר לדעת את e , שהיא הסטייה מקו הרגרסיה. נגדיר זאת באופן הבא:

* עבור X_i , הערך הצפוי של המשתנה המוסבר (\hat{Y}_i) המתקבל לפי הרגרסיה הוא $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$.

* הסטייה של התצפית (Y_i) מהערך הצפוי לפי הרגרסיה (\hat{Y}_i) היא: $e_i = Y_i - \hat{Y}_i$



האומדים של הרגרסיה $(\hat{\alpha}, \hat{\beta})$:

שיטת האמידה של α ושל β נקראת שיטת הריבועים הפחותים

Ordinary Least Squares (OLS)

השאלה הנשאלת בשיטת אמידה זו היא: איזה $\hat{\alpha}$ ו- $\hat{\beta}$ יביאו למינימום את סכום ריבועי טעויות האמידה.

$$\min_{\hat{\alpha}, \hat{\beta}} \sum e_t^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum (y_t - \hat{y}_t)^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum [y_t - (\hat{\alpha} + \hat{\beta}x_t)]^2 = ?$$

מתוך גזירת הפונקציה הזו מתקבלים האומדים $\hat{\alpha}$ ו- $\hat{\beta}$:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{SXY}{SXX} = \frac{COV(X, Y)}{V(X)} = r \frac{S_Y}{S_X}$$

? על סמך הדוגמא הנ"ל חשבו את:

1. מקדם המתאם בין גודל הדירה למחיר הדירה. מה משמעותו?
2. קו הרגרסיה לניבוי מחיר הדירה באמצעות גודל הדירה ופרשו את משמעות המקדמים.
3. המחיר החזוי על פי קו הרגרסיה של דירה בגודל 100 מ"ר.

מבחני המובהקות

השערות: $H_0: \beta = 0$

$H_1: \beta \neq 0$

ברגרסיה פשוטה בה יש לנו רק מנבא אחד: ניתן לבצע מבחן F למובהקות משוואת הרגרסיה או מבחן T למובהקות מקדם הרגרסיה (הביטא).

משמעות דחיית השערת האפס: משוואת הרגרסיה מובהקת, מקדם הרגרסיה מובהק, הקשר בין X ל-Y מובהק.

ולחיפך- אם השערת האפס לא נדחית: אין הוכחה לקשר בין המשתנים X ו-Y, משוואת הרגרסיה איננה מובהקת וכך גם מקדם הרגרסיה.

אמידת σ^2 שונות הטעויות:

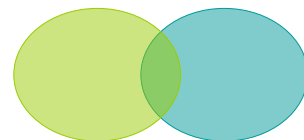
$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{(1-r^2)SST}{n-2}$$

מבחן F

מבחן זה נעשה על מנת לבדוק האם משוואת הרגרסיה מובהקת.

מבחן F מתבסס על פירוק סכום הריבועים:

$$\underbrace{SST}_{S_Y^2} = \underbrace{SSR}_{r^2 S_Y^2} + \underbrace{SSE}_{(1-r^2) S_Y^2}$$



טבלת ניתוח שונות (טבלת ANOVA)

מקור	סכום ריבועים SS	דרגות חופש d.f.	ממוצע סכום ריבועים MS=SS/d.f.	F
מודל הרגרסיה	SSR	1	MSR=SSR/1	MSR/MSE
שאריות	SSE	n-2	MSE=SSE/n-2	
סה"כ	SST	n-1		

כלל הכרעה:

אם $F_{st} > F_c(1, n-2)$ נדחה את השערת האפס.

? בהמשך לדוגמא הנ"ל:

בצעו מבחן F לבדיקת הקשר בין גודל הדירה למחירה ברמת מובהקות של 1%.

הערה: ניתן גם לשאול- האם משוואת הרגרסיה לניבוי מחיר הדירה על סמך גודלה מובהקת באוכלוסיה ברמת מובהקות של 1%?

מבחן t

מבחן זה נעשה על מנת לבדוק האם מקדם הרגרסיה מובהק.

סטטיסטי המבחן:

$$t_{st} = \frac{\hat{\beta}_1 - \beta_{1,0}}{s.e.(\hat{\beta}_1)} \sim t_{c(n-2)}$$
$$s.e.(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SXX}}$$

אם השערת האפס מתייחסת ל- $\beta=0$ (בדר"כ):

$$t_{stt} = \frac{r^2 \sqrt{n-2}}{\sqrt{1-r^2}}$$

כלל הכרעה:

השערה דו צדדית $H_1 : \beta_1 \neq \beta_{1,0}$	השערה חד צדדית שמאלית $H_1 : \beta_1 < \beta_{1,0}$	השערה חד צדדית ימנית $H_1 : \beta_1 > \beta_{1,0}$	
$t_{\text{statistic}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{s.e.}(\hat{\beta}_1)} = \frac{r^2 \sqrt{n-2}}{\sqrt{1-r^2}}$ $s.e.(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SXX}}$			סטטיסטי המבחן
$ t_{\text{statistic}} \geq t_{n-2, 1-\alpha/2}$	$t_{\text{statistic}} \leq -t_{n-2, 1-\alpha}$	$t_{\text{statistic}} \geq t_{n-2, 1-\alpha}$	אזור דחייה
$2 * P(t_{n-2} > t_{\text{statistic}})$	$P(t_{n-2} > t_{\text{statistic}})$	$P(t_{n-2} > t_{\text{statistic}})$	P-VALUE

? בהמשך לדוגמא הנ"ל:

1. בצעו מבחן t למובהקות מקדם הרגרסיה ברמת מובהקות של 1%.

אפשר גם לבקש: בצעו מבחן t לבדיקת הקשר בין גודל הדירה למחירה.

2. בדקו את הטענה כי עליה במ"ר אחד תעלה את מחיר הדירה ביותר מ-2000\$.

3. מהו ה-pvalue של מובהקות הקשר בין גודל הדירה למחירה. מה משמעותו?

? חשבו את סטטיסטי המבחן F על סמך סטטיסטי המבחן t שקיבלתם. מה ה-pvalue של

מבחן F?

רווח סמך לאמידת β :

$$p(\text{גבול תחתון} \leq \beta \leq \text{גבול עליון}) = 1 - \alpha$$

$$\hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} \cdot s.e.(\hat{\beta}_1)$$

? חשבו רב"ס לאמידת מקדם הרגרסיה ברמת סמך של 0.99. השוו עם תוצאות מבחן t.

מדד טיב ההתאמה R^2 :

מדד שנותן את פרופורציית השונות המוסברת. כמה מהשונות של Y מוסברת על ידי השונות של X:

$$0 \leq R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \leq 1 \quad (\text{X מסביר את כל השונות של Y})$$

נרצה פרופורציית שונות מוסברת קרובה ככל האפשר ל-1.

אחוז השונות המוסברת: $R^2 \cdot 100$

? חשבו את אחוז השונות המוסברת של מחיר הדירה על ידי גודלה.

תרגול מסכם

בפיצויית "שלמה המלך" חושדים כי מספר הלקוחות המבקרים בפיצויה תלוי במחיר המכירה של הבירה במקום. לשם בדיקת הנושא ערכו ניסוי בו בכל שבוע שינו את מחיר הבירה במקום ומנו את מספר הלקוחות שהגיעו במשך אותו שבוע. משך הניסוי 7 שבועות עוקבים. להלן נתוני הניסוי:

שבוע	שבוע	שבוע	שבוע	שבוע	שבוע	שבוע	
9	10	11	12	13	14	15	מחיר הבירה
164	155	150	150	148	145	143	כמות הלקוחות

- א. אמדו את מודל הרגרסיה ע"י חישוב מקדמי הרגרסיה
- ב. חשבו את מקדם המתאם r_{xy}
- ג. אמדו את השונות של שאריות המודל
- ד. חשבו את אחוז השונות המוסברת. מה משמעותה?
- ה. בצעו חיזוי לכמות הלקוחות אם מחיר הבירה יהיה 16 ₪. האם להערכתכם ניתן להיסתמך על חיזוי זה?
- ו. בצעו מבחן F לבדיקה האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצויה ברמת מובהקות 5%.
- ז. בצעו מבחן t לבדיקה האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצויה ברמת מובהקות 5%. השוו את התוצאות.
- ח. אמדו את מקדם הרגרסיה ברמת סמך של 0.95. השוו את התוצאה עם הסעיף הקודם.

רגרסיה פשוטה-פלטי SPSS

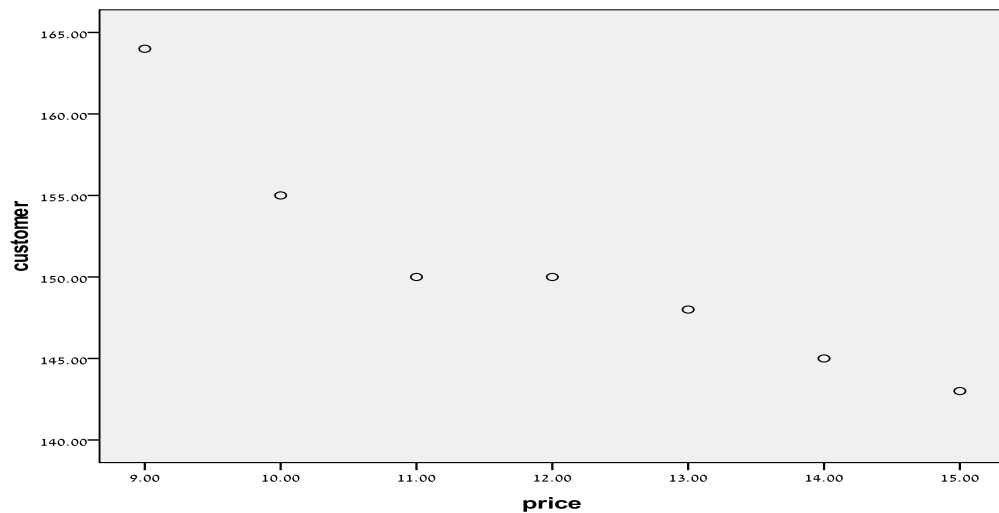
הבא נראה כיצד לקרוא פלטי SPSS ברגרסיה פשוטה.

על סמך הנתונים של שאלה מס' 1 :

שבוע 7	שבוע 6	שבוע 5	שבוע 4	שבוע 3	שבוע 2	שבוע 1	
9	10	11	12	13	14	15	מחיר הבירה (ש)
164	155	150	150	148	145	143	כמות הלקוחות

התקבלו הפלטים הבאים:

(1) דיאגרמת הפיזור (scatter plot):



(2) סטטיסטיקה תיאורית (descriptive statistics):

Descriptive Statistics

	Mean	Std. Deviation	N
customer	150.7143	7.01699	7
Price	12.0000	2.16025	7

(3) פלט מקדם המתאם (correlations):

		customer	Price
Pearson Correlation	customer	1.000	-.935
	price	-.935	1.000
Sig. (1-tailed)	customer	.	.001
	price	.001	.
N	customer	7	7
	price	7	7

(4) פלט model summary:

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935 ^a	.873	.848	2.73470

a. Predictors: (Constant), price

b. Dependent Variable: customer

(5) פלט ניתוח שונות (ANOVA):

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	258.036	1	258.036	34.503	.002 ^a
	Residual	37.393	5	7.479		
	Total	295.429	6			

a. Predictors: (Constant), price

b. Dependent Variable: customer

6) פלט מקדמי הרגרסיה (coefficients):

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	187.143	6.287		29.765	.000
	Price	-3.036	.517	-.935	-5.874	.002

a. Dependent Variable: customer

? על סמך הפלטים הנתונים:

- א. מהו מודל הרגרסיה שנאמד?
- ב. מהו מקדם המתאם r_{xy} ?
- ג. מהי השונות של שאריות המודל?
- ד. מהו אחוז השונות המוסברת?
- ה. על פי מבחן F: האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצוצייה ברמת מובהקות 5%?
- ו. על פי מבחן t: האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצוצייה ברמת מובהקות 5%? השוו את התוצאות.
- ז. מה ה-pvalue של המבחנים הסטטיסטיים? מה משמעותו?
- ח. בדקו האם קיים קשר חיובי מובהק בין המשתנים ברמת מובהקות 5%?

רגרסיה מרובה

ניבוי המשתנה התלוי באמצעות יותר ממשתנה ב"ת אחד.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

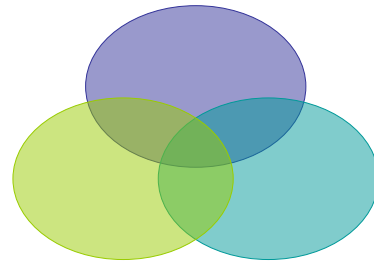
המודל באוכלוסיה:

מקדמי מודל הרגרסיה המרובה:

α = חותך אחד שמשמעותו: הציון המנובא כאשר כל המשתנים הב"ת=0.

$\beta_1 \dots \beta_j$ = מקדמי השיפוע. מס' הבטות = למספר המשתנים הב"ת במודל.

משמעות מקדם השיפוע β_j : ההשפעה הייחודית של המשתנה הב"ת המסוים לניבוי המשתנה התלוי, בניכוי השפעתם של כל יתר המשתנים הב"ת האחרים המצויים במשוואת הרגרסיה.



אמידת מודל הרגרסיה המרובה:

ברגרסיה מרובה, כמו ברגרסיה פשוטה, שיטת האמידה הטובה ביותר היא שיטת הריבועים הפחותים. כלומר, נרצה להביא את סכום הטעויות בניבוי למינימום.

מפיתרון פונקצית הריבועים הפחותים נקבל את אומדי הרגרסיה: $\hat{\alpha}, \hat{\beta}_1 \dots \hat{\beta}_j$

מבחני מובהקות:

(1) מבחן F למובהקות הרגרסיה

בדיקה האם קיים קשר ליניארי בין המשתנה התלוי Y לבין לפחות אחד מהמשתנים המסבירים.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

השערות הן:

$H_1 : \text{Not } H_0 = \text{at least one of the } \beta\text{'s is not 0}$

טבלת ניתוח שונות (ANOVA)

מקור	סכום ריבועים SS	דרגות חופש d.f.	ממוצע סכום ריבועים MS=SS/d.f.	$F_{st} \sim F_{k,n-k-1}$
מודל הרגרסיה	SSR	k	MSR=SSR/ k	$F_{st} = \text{MSR}/\text{MSE}$
שאריות	SSE	$n-k-1$	MSE=SSE/ $(n-k-1)$	
סה"כ	TSS	$n-1$		

$$F_{st} = \frac{\text{MSR}}{\text{MSE}} \quad \text{סטטיסטי המבחן:}$$

כלל הכרעה: נדחה את H_0 אם $F_{st} \geq F_{k,n-k-1}^{1-\alpha}$.

חישוב סכומי הריבועים:

$$\begin{aligned} TSS &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 \\ SSR &= R^2 \cdot TSS \\ SSE &= (1 - R^2)TSS \end{aligned}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

פרופורציית השונות המוסברת R^2 :

ברגרסיה מרובה אומד זה לפרופורציית השונות המוסברת הוא בעייתי שכן הוא מושפע ממספר המשתנים ה"ב"ת במודל. אומד זה יכול רק לגדול בהוספת משתנים ב"ת למודל ולכן לא ייתן לנו אינדיקציה האם כדאי היה לי להוסיף אותם למודל או לא.

האומד המתוקן לפרופורציית השונות המוסברת $AdjR^2$:

$$\bar{R}^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

בניגוד ל- R^2 לוקח בחשבון את מספר המשתנים ה"ב"ת במודל. יכול שלא לגדול ואף לקטון בהוספת משתנה ב"ת שלא תורם תרומה משמעותית לניבוי.

(2) מבחן t למובהקות משתנה ב"ת יחיד:

השערות:

$$H_0 : \beta_j = 0$$

$$H_1 : \text{else}$$

סטטיסטי המבחן וכלל הכרעת השערת האפס:

$$\left| T = \frac{\text{אמון מקדם}}{\text{סטיית תקן מקדם}} \right| > t_{1-\frac{\alpha}{2}}^{(n-k-1)} \quad \text{for } n < 30$$

רווח בר סמך לאמידת ה- β_j :

$$\hat{\beta}_j \pm t_{n-k-1, 1-\alpha/2} \text{ s.e.}(\hat{\beta}_j)$$

(3) מבחן F חלקי (partial F):

בודק את ההשערה שתוספת של משתנה אחד או קבוצה של משתנים מוסיפה תוספת מובהקת לניבוי המשתנה התלוי מעבר למשתנים אחרים שקיימים כבר במודל.

השערות:

$$p < k \quad \begin{array}{l} H_0 : \beta_1 = \beta_2 = \dots = \beta_p \\ H_1 : \text{אחרת} \end{array}$$

ביצוע המבחן:

מריצים שתי רגרסיות:

(1) UR - המודל המלא-רגרסיה עם כל המשתנים הב"ת (K)

(2) R - המודל החלקי-רגרסיה תחת H0 (K-P)

סטטיסטי המבחן:

$$F = \frac{R_{UR}^2 - R_R^2 / p}{1 - R_{UR}^2 / n - k - 1} = \frac{SSR_{UR} - SSR_R / p}{1 - SSR_{UR} / n - k - 1} = \frac{SSE_R - SSE_{UR} / p}{1 - SSE_{UR} / n - k - 1}$$

כלל הכרעה:

$$F > f_{1-\alpha}^{p, n-k-1}$$

לדוגמא: נתונים 4 משתנים ב"ת לניבוי משתנה תלוי מסויים. רוצים לבדוק האם משתנה X1 ו-X2 מוסיפים תוספת משמעותית לניבוי Y מעבר למשתנים X3 ו-X4.

בהרצת רגרסיה עם כל המשתנים הב"ת התקבל $R^2 = 0.982$

בהרצת רגרסיה עם משתנים X3 ו-X4 בלבד התקבל $R^2 = 0.935$

השערות:

$$H0: \beta_1 = \beta_2 = 0$$

H1: אחרת

חישוב סטטיסטי המבחן:

$$F = \frac{R_{UR}^2 - R_R^2 / p}{1 - R_{UR}^2 / n - k - 1} = \frac{0.982 - 0.935 / 2}{1 - 0.982 / 12 - 4 - 1} = \frac{0.0235}{0.00257} = 9.144$$

כלל הכרעה:

$$F = 9.144 > F_{0.95}^{2,9} = 4.257$$

לכן יש סיבה מספקת לדחות את H0 ברמת מובהקות של 0.05.

מסקנה: המשתנים X1 ו-X2 מוסיפים תוספת מובהקת לניבוי של Y מעבר ליתר המשתנים ה"ת במשוואה (X3 ו-X4).

קשר בין מבחן F חלקי למבחן t :

קיים קשר בין מבחן F למובהקות תוספת משתנה ב"ת יחיד למבחן t למובהקות אותו משתנה :

$$F_{1-\alpha}^{1, n-k-1} = t_{1-\frac{\alpha}{2}, n-k-1}^2$$

$$pvalue = pvalue$$

תירגול

לצורך בדיקת ההשערה שקיים קשר בין מספר המוניות בעיר באר שבע (y) לבין מספר התושבים בעיר באלפים (x1) ומספר הרכבים הפרטיים באלפים (x2). הוחלט לבנות מודל רגרסיה מהצורה: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, על סמך הנתונים הבאים:

$$MSE=119.789, \sum_i y_i^2 = 338657, \sum_i y_i = 1673$$

א. ע"ס הנתונים הנ"ל, השלימו את טבלת ניתוח השונות הבאה. איזו השערה ניתן לבדוק באמצעותה? כתוב את ההשערה ובחן אותה.

SOURCE	SS	DF	MS	F
Regression				
Error				
Total		8		

ב. חשבו את מדד טיב ההתאמה. הסבר את משמעותו.

ג. נתונה טבלת המקדמים (החלקית) הבאה:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-511.727	114.9476				
X 1	9.208785		3.732167			
X 2	-8.79921	4.420456				

1. רשמו את האומדן למשוואת הרגרסיה ופרשו את מקדמיה.
2. בחנו את ההשערה כי קיים קשר בין מספר הרכבים הפרטיים לבין מספר המוניות ברמת מובהקות של 5%.
3. בנו רווח סמך למקדם של מספר התושבים בעיר.
4. ענה ללא חישוב (על סמך הסעיפים הקודמים) - האם קיים קשר בין מספר התושבים לבין מספר המוניות ברמת מובהקות של 5%?
5. מהי תחזית מס' המוניות בבאר שבע עבור 100,000 תושבים ו-52,000 מכוניות פרטיות?
6. האם ניתן לסמוך על תחזית זאת?
7. חשב את סטטיסטי F חלקי של מס' הרכבים הפרטיים. האם מובהק (ענה ללא חישוב).

חישוב מובהקות התוספת (F חלקי) של משתנה ב"ת מסוים על פני האחרים:

במקרה של מולטיקוליניאריות במודל (מתאם חזק בין משתנים ב"ת), בכדי לדעת איזה משתנה ב"ת יש להוציא, ניתן לבחון את התוספת לניבוי של המשתנה ה"חשוד" על פני האחרים. אם היא איננה מובהקת, זוהי אינדיקציה שיש להוציא מהמודל.

במקרה שלנו נבחן את התוספת לניבוי של X4 על פני המשתנים הב"ת האחרים:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.991 ^a	.982	.974	2.44601

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.991 ^a	.982	.974	2.44601

a. Predictors: (Constant), X4, X3, X1, X2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.991 ^a	.982	.976	2.31206

a. Predictors: (Constant), X3, X2, X1

$$F = \frac{(R_{UR}^2 - R_R^2) / P}{(1 - R_{UR}^2) / n - k - 1} = \frac{(0.982 - 0.982) / 1}{(1 - 0.982) / 13 - 4 - 1} = 0$$

מסקנה: X4 לא מוסיף תוספת מובהקת למודל.

תרגול מסכם:

מעוניינים למצוא קשר בין מחיר הדירה (ב-\$) לבין ארבעה משתנים מסבירים: (1) שטח הדירה ו- (2) גודל שטח האמבטיה (ב-Sqft) וכן (3) מרחק הדירה מהים ו- (4) מהאוניברסיטה (במיילים).

לשם כך נדגמו מספר דירות והריצו רגרסיה אשר בה המשתנה המוסבר הוא מחיר הדירה. להלן פלט הרגרסיה שהתקבל:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.952 ^a	○	○	○

a. Predictors: (Constant), Sea_Dist, Apartment, Bath

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression					.000 ^a
	Residual					
	Total	1940484.615	25			

a. Predictors: (Constant), Univ_Dist, Bath, Sea_Dist, Apartment

b. Dependent Variable: Price

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-265.514	146.673		-1.810	.085
	Apartment		.449	.722	6.572	
	Bath	4.256		.297	2.687	.014
	Sea_Dist	-32.114	11.090	-.223		.009
	Univ_Dist	11.746	9.439	.095	1.244	.227

a. Dependent Variable: Price

ענה על הסעיפים הבאים:

- א. מלאו את התאים החסרים בטבלה (אם לא ניתן למלא את כל התאים החסרים באופן מלא נמקו באופן מפורש מדוע לא ניתן).
- ב. כתבו את האומדן למשוואת מחיר הדירה בצורה מפורשת על סמך הפלט הנ"ל. פרשו את מקדמי הרגרסיה.
- ג. בדקו האם ארבעת הגורמים ביחד אכן מסבירים את מחיר הדירה. הסברו את המסקנה שהגעתם אליה. השתמשו ברמת מובהקות 5%.
- ד. הסברו מהו ערך ה-Pvalue ומה ניתן להסיק ממנו לגבי המשתנים המסבירים?
- ה. בנו רווח סמך למקדם גודל שטח האמבטיה. השתמשו ברמת מובהקות של 2%.
- ו. ברמת מובהקות של 5% יש לבדוק האם המרחק מהאוניברסיטה אכן משפיע על מחיר הדירה.
- ז. האם במודל הרגרסיה הנוכחי ניתן לוותר על גורם המרחק מהים? השתמשו ברמת מובהקות 1%.
- ח. בדקו את ההשערה כי קיים קשר חיובי בין גודל הדירה למחירה ברמת מובהקות של 5%.

ט. נתונה מטריצת מקדמי המתאם הבאה:

	$X1$	$X2$	$X3$	$X4$
$X1$	1			
$X2$	0.228579	1		
$X3$	-0.22413	-0.13924	1	
$X4$	-0.24545	-0.97295	0.029537	1

מה ניתן ללמוד ממנה ומה משמעותה לגבי המודל?

י. האם משתנים $X2$ ו- $X4$ מוסיפים תוספת משמעותית לניבוי? אם לא ניתן לענות על השאלה, ציין מדוע.

יא. מה יהיו תוצאות מבחן F לבדיקת התוספת לניבוי של המרחק מהאוניברסיטה על פני המשתנים האחרים (ענה ללא חישוב)

משתני דמי

הנושא של משתני דמי מטפל בהכנסת משתנים ב"ת איכותיים למודל הרגרסיה.

עד כה כל המשתנים הב"ת שהכנסנו למודל היו כמותיים, כלומר קיבלו ערכים מספריים.

למשל, נניח שאנו סבורים שמס' שנות הלימוד של אדם משפיעות על שכרו:

$$W_t = \text{השכר (המשתנה התלוי)}$$

$$S_t = \text{שנות לימוד (המשתנה הב"ת)}$$

משוואת הרגרסיה:

$$W_t = \alpha + \beta \cdot S_t$$

במקרה זה המשתנה המסביר (כמו גם המוסבר) הוא כמותי.

נניח שאנו סבורים שגם משתנה המגדר משפיע על השכר. משתנה זה איננו כמותי כמו שנות לימוד אלא איכותי שכן הוא לא מקבל ערכים מספריים אלא ערכים קטגוריאליים כ"גבר" או "אישה".

נשאלת השאלה כיצד נכניס אותו לתוך משוואת הרגרסיה?

נגדיר משתנה D שיקבל את הערך 0 אם מדובר ב"אישה" ואת הערך 1 אם מדובר ב"גבר".

משתנה כזה נקרא משתנה דמי (dummy variable).

לעומת משתנה רגיל ש"פועל" תמיד, משתנה זה "יפעל" רק אם מדובר בגבר.

ניתן להכניס את משתנה הדמי למודל בשלושה אופנים שונים:

(1) משתנה דמי לחותך- המין משפיע על השכר ההתחלתי בלבד

(2) משתנה דמי לשיפוע- המין משפיע על התוספת לשכר בגין שנות הלימוד

(3) משתנה דמי לכל הפונקציה- המין משפיע גם על החותך וגם על השיפוע

(1) משתנה דמי לחותך

המין משפיע על השכר ההתחלתי בלבד.

$$W_t = \alpha_0 + \alpha_1 D + \beta \cdot S_t + u_t$$

החותך מייצג כאן את השכר ההתחלתי.

שכר ההתחלתי של אישה: α_0

שכר התחלתי של גבר: $\alpha_0 + \alpha_1$

הבדל בשכר בין נשים וגברים: α_1 (הפרש בין החותכים)

בדיקת השערות על משתנה הדמי: מבחן t למובהקות הפרש החותכים: $H_0: \alpha_1 = 0$

** השיפוע מייצג את התוספת לשכר כפונקציה של מס' שנות הלימוד והוא זהה עבור נשים וגברים.

? על בסיס מדגם של 50 איש העובדים בחברה מסוימת התקבלו התוצאות הבאות:

$$W_t = 5500 + 1043 \cdot D + 119 \cdot S_t$$

(S.E) (134) (56) (24)

המספרים בסוגריים הם טעויות התקן של מבחני המובהקות לפרמטרים.

- מהו השכר ההתחלתי של גבר בעל 12 שנות לימוד?
- מה ההבדל בשכר ההתחלתי בין גברים לנשים?
- האם הבדל זה מובהק באוכלוסיה?
- בדקו את הטענה כי השכר ההתחלתי של גברים גבוה ביותר מ-500 מזה של נשים.
- בדקו את הטענה שהשכר ההתחלתי של נשים נמוך ב-600 מזה של גברים.

(2) משתנה דמי לשיפוע

המגדר משפיע על התוספת לשכר בגין שנות הלימוד.

$$W_t = \alpha + \beta_0 S_t + \beta_1 DS_t + u_t$$

השיפוע מייצג כאן את התוספת לשכר בגין שנות לימוד.

אצל אישה: התוספת לשכר בגין שנות לימוד- β_0

אצל גבר: התוספת לשכר בגין שנות לימוד- $\beta_0 + \beta_1$

הבדל בין גברים לנשים בתוספת לשכר בגין שנות הלימוד: β_1 (הפרש השיפועים)

בדיקת השערות על משתנה הדמי: מבחן t למובהקות הפרש השיפועים: $H_0: \beta_1 = 0$

** החותך, המייצג את השכר ההתחלתי, יהיה זהה עבור גברים ונשים.

? על בסיס אותו מדגם, ביקש החוקר לדעת האם קיים הבדל מובהק בין גברים לנשים

בתוספת לשכר בגין שנות הלימוד. תוצאות האמידה נתונות להלן:

$$W_t = 5000 + 110 \cdot S_t + 120 \cdot D \cdot S_t + u_t$$

$$(68) \quad (23) \quad (25)$$

בדוק את ההשערה.

? חברה מסוימת מתמרצת עובדים ונותנת להם תגמול לפי מספר השעות הנוספות שעבדו בחודש.

נגדיר $Y = \text{סה"כ התגמול שמקבל עובד שעבד } X \text{ שעות נוספות בחודש.}$

$D=1$ אם $10 < X$ ו $D=0$ אחרת.

איזו משוואת רגרסיה יש להריץ בכדי לאמוד את סה"כ התגמול בתנאים הבאים:

1. מי שעבד עד וכולל 10 שעות נוספות בחודש תגמולו יהיה קבוע וזהה לכל השעות הנוספות ואילו מי שעבד מעבר ל-10 שעות נוספות מקבל תגמול שונה לכל שעה נוספת מעבר ל-10 הראשונות.

2. עובדים שעבדו יותר מ-10 שעות נוספות בחודש מקבלים סכום קבוע ללא תלות במספר השעות הנוספות שעבדו.

3.

(3) משתנה דמי לכל הפונקציה

המין משפיע גם על החותך וגם על השיפוע. הווה אומר, גם על השכר ההתחלתי וגם על התוספת לשכר בגין שנות הלימוד.

$$\text{המודל: } W_t = \alpha_0 + \alpha_1 D + \beta_0 S_t + \beta_1 DS_t + u_t$$

השכר ההתחלתי של אישה: α_0

השכר ההתחלתי של גבר: $\alpha_0 + \alpha_1$

הבדל בשכר ההתחלתי בין המינים: α_1 (הבדל בחותכים)

אצל אישה- התוספת לשכר בגין שנות הלימוד: β_0

אצל גבר- התוספת לשכר בגין שנות הלימוד: $\beta_0 + \beta_1$

הבדל בין המינים בתוספת לשכר בגין שנות הלימוד: β_1 (הבדל בשיפועים)

בדיקת השערות למשתני הדמי:

$$H0: \alpha_1 = \beta_1 = 0$$

באמצעות מבחן WALS יש לבדוק:

H1: לפחות אחד הפרמטרים שונה מ-0

אם דוחים את השערת האפס, יש לבצע מבחני t עבור כל אחד מהפרמטרים בנפרד:

$$H0: \beta_1 = 0 \quad \text{vs} \quad H0: \alpha_1 = 0$$

? חוקר רצה לבדוק את הטענה שסוג הכביש משפיע על מס' תאונות הדרכים בקטעי כביש בינעירוניים, בהינתן נפח התנועה.

החוקר בדק האם הפונקציה של מס' התאונות בהינתן נפח התנועה, שונה בין כבישים מהירים לבין כבישים שאינם מהירים. לשם כך אסף החוקר 754 תצפיות ואמד את המשוואות הבאות:

$$NUM_t = \gamma_3 + \delta_3 \cdot AVGD_t + \varepsilon_{3t} \quad (1)$$

$$NUM_t = \alpha + \beta_1 \cdot TYPE_t + \beta_2 \cdot AVGD_t + \beta_3 \cdot (AVGD \cdot TYPE)_t + U_t \quad (2)$$

כאשר: NUM_t = מס' תאונות הדרכים הקטלניות בקטע כביש t בשנה

$AVGD_t$ = נפח התנועה בקטע כביש t ליום באלפים

$TYPE_t$ = משתנה דמי המקבל את הערך 1 כאשר הכביש מהיר, ו-0 כאשר הכביש לא מהיר.

תוצאות אמידת המשוואות מוצגות להלן:

$$NUM_t = 0.739 + 0.0233 \cdot AVGD_t \quad (1)$$

$$NUM_t = 0.14978 + 1.40311 \cdot TYPE_t + 0.002877 \cdot AVGD_t - 0.008 \cdot (AVGD \cdot TYPE)_t \quad (2)$$

$$ESS = 20963 \quad Pt_{\hat{\alpha}} = 0.0019; Pt_{\hat{\beta}} = 0.0001 \quad (1)$$

$$Pt_{\hat{\alpha}} = 0.6534; Pt_{\hat{\beta}_1} = 0.0067; Pt_{\hat{\beta}_2} = 0.0001; Pt_{\hat{\beta}_3} = 0.1283 \quad ESS = 20759 \quad (2)$$

(1) בדקו את טענת החוקר.

(2) מהו האומדן הנקודתי למס' התאונות בכביש מהיר כאשר נפח התנועה עומד על 4 אלפי מכוניות ליום בקטע הכביש האמור?

הועלתה הטענה כי המקדם להשפעה של נפח התנועה בדרכים מהירות הינו כפול מזה שבדרכים לא-מהירות.

(3) מהי השערת האפס לבדיקת הטענה?

(4) מהי הרגרסיה "תחת" H_0 למבחן WALT ?

סיכום ביניים:

משתנה דמי לכול הפונקציה	משתנה דמי לשיפוע	משתנה דמי לחותך	
$Y_t = \alpha_0 + \alpha_1 D + \beta_0 X_t + \beta_1 DX_t + u_t$	$Y_t = \alpha + \beta_0 X_t + \beta_1 DX_t + u_t$	$Y_t = \alpha_0 + \alpha_1 D + \beta \cdot X_t + u_t$	המודל
קיים הבדל בין הקטגוריות במשוואת הרגרסיה כולה (בחותרך ובשיפוע).	קיים הבדל בין הקטגוריות בתוספת ל- Y בגין X (בשיפוע).	קיים הבדל בין הקטגוריות ב- Y ההתחלתי (בחותרך).	ההשערה במילים
מבחן WALT להפרש בין הפונקציות (החותכים והשיפועים): $H0: \alpha_1 = \beta_1 = 0$ אם דוחים את $H0$ יש לברר את מקור ההבדל באמצעות מבחני t (אפשרי רק ב-WALT): $H0: \alpha_1 = 0$ $H0: \beta_1 = 0$	מבחן t להפרש השיפועים: $H0: \beta_1 = 0$	מבחן t להפרש החותכים: $H0: \alpha_1 = 0$	בדיקת ההשערה

משתני דמי אם המשתנה האיכותי יכול לקבל יותר משני ערכים

כאשר המשתנה האיכותי כלל שני ערכים בלבד (למשל, מגדר: גבר, אישה) הסתפקנו במשתנה דמי אחד.

במקרים רבים המשתנה האיכותי כולל יותר משני ערכים/קטגוריות. במקרה כזה נגדיר מס' משתני דמי כמספר הקטגוריות פחות אחד.

למשל, את המשתנה האיכותי של עונות השנה הכולל 4 ערכים: אביב, קיץ, סתיו, חורף נייצג באמצעות 3 משתני דמי:

D_1 יקבל את הערך 1 אם מדובר באביב ו-0 אחרת.

D_2 יקבל את הערך 1 אם מדובר בקיץ ו-0 אחרת.

D_3 יקבל את הערך 1 אם מדובר בסתיו ו-0 אחרת.

אם מדובר בחורף אז כל משתני הדמי יקבלו את הערך 0 ולכן החורף היא קבוצת הייחוס.

נניח שאנו רוצים לבדוק עונתיות במחירי הירקות:

$$V_t = \text{מדד מחירי הירקות}$$

$$p_t = \text{מדד המחירים לצרכן}$$

(1) משתני דמי לחותך

הטענה: יש הבדל בין עונות השנה במחיר ההתחלתי של הירקות

$$\text{המודל: } V_t = \alpha_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \beta \cdot P_t + u_t$$

כל עליה של יחידה אחת במדד המחירים לצרכן תעלה את מחירי הירקות ב- β . למחיר

זה יתווסף α_0 בחורף, $\alpha_0 + \alpha_1$ באביב, $\alpha_0 + \alpha_2$ בקיץ ו- $\alpha_0 + \alpha_3$ בסתיו.

ניתן לראות כי:

α_0 : החותך בקטגוריה שהושמטה $\alpha_0 + \alpha_i$: החותך בקטגוריה i.

בדיקת השערות:

השערות:

$$H0: \alpha_1 = \alpha_2 = \alpha_3 = 0$$

$$H1: OTHERWISE$$

המבחן הסטטיסטי :

מבחן WALT:

$$V_t = \alpha_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \beta \cdot P_t + u_t \quad (U)$$

$$V_t = \alpha + \beta \cdot P_t + u_t \quad (R)$$

**שימו לב שהחותך במשוואה המוגבלת איננו α_0 שכן המשתנה המסביר של עונות השנה ירד.

אם נדחה את HO במבחן הסטטיסטי של הסעיף הקודם, יש לבדוק מה מקור ההבדל בין

החותכים על ידי מבחני t:

1. האם יש הבדל במחיר ההתחלתי של הירקות בין האביב לחורף: $H0: \alpha_1 = 0$

2. האם יש הבדל במחיר ההתחלתי של הירקות בין הקיץ לחורף: $H0: \alpha_2 = 0$

3. האם יש הבדל במחיר ההתחלתי של הירקות בין הסתיו לחורף: $H0: \alpha_3 = 0$

? א. הועלתה הטענה כי יש הבדל במחיר ההתחלתי בין האביב לקיץ.

1. מהי השערת האפס לבדיקת הטענה?

2. פרטו שני מבחנים סטטיסטיים בעזרתם ניתן לבדוק את הטענה.

3.

ב. הועלתה הטענה כי יש רק שתי עונות המשפיעות על מחיר הירקות ההתחלתי: קיץ+אביב, חורף+סתיו.

1. מהי השערת האפס לבדיקת הטענה?

2. מהו המבחן הסטטיסטי המתאים? פרטו.

? כלכלן הציע לאמוד את המודל הבא: $ta25_t = \alpha_0 + \alpha_1 Q1_t + \alpha_2 Q2_t + \alpha_3 Q3_t + \alpha_4 Q4_t + u_t$

כאשר :

Ta25 - תשואת מדד המניות ת"א 25.

Q_i - משתנה דמי המקבל ערך 1 עבור רבעון i ו-0 אחרת.

מי מהטענות הבאות נכונה (יש רק אחת):

1. לא ניתן לאמוד את המודל כיוון שאין שום משתנה מסביר אלא רק משתני דמי.

2. לא ניתן לאמוד את המודל יש בעיית מולטיקוליניאריות מלאה.

3. אין שום בעיה לאמוד את המודל.

4. כל התשובות האחרות שגויות.

(2) משתני דמי לשיפוע

הטענה: יש הבדל בין עונות השנה בתוספת למחיר הירקות בגין המחיר לצרכן

המודל: $V_t = \alpha + \beta_0 P_t + \beta_1 (D_{1i} P_t) + \beta_2 (D_{2i} P_t) + \beta_3 (D_{3i} P_t) + u_t$

המחיר ההתחלתי של הירקות שווה בין עונות השנה (α) אולם כל עליה של יחידה אחת

במדד המחירים לצרכן תעלה את מחירי הירקות ב:

β_0 בחורף, $\beta_0 + \beta_1$ באביב, $\beta_0 + \beta_2$ בקיץ ו- $\beta_0 + \beta_3$ בסתיו.

ניתן לראות כי:

β_0 : השיפוע בקטגוריה שהושמטה

$\beta_0 + \beta_i$: השיפוע בקטגוריה i.

בדיקת השערות:

השערות:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$H_1: \text{OTHERWISE}$

המבחן הסטטיסטי:

מבחן WALT:

$$V_t = \alpha + \beta_0 P_t + \beta_1 (D_{1i} P_t) + \beta_2 (D_{2i} P_t) + \beta_3 (D_{3i} P_t) + u_t \quad (U)$$

$$V_t = \alpha + \beta \cdot P_t + u_t \quad (R)$$

**שימו לב שהשיפוע במשוואה המוגבלת איננו β_0 שכן המשתנה המסביר של עונות השנה ירד.

אם נדחה את H_0 במבחן הסטטיסטי של הסעיף הקודם, יש לבדוק מה מקור ההבדל בין השיפועים על ידי מבחני t.

(3) משתני דמי לכל הפונקציה

הטענה: יש הבדל בין עונות השנה בפונקצית הרגרסיה לניבוי מחיר הירקות באמצעות המחיר לצרכן.

$$V_t = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta_0 P_t + \beta_1 (D_{1i} P_t) + \beta_2 (D_{2i} P_t) + \beta_3 (D_{3i} P_t) + u_t \quad \text{המודל:}$$

בדיקת השערות:

השערות:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = \beta_3 = 0$$

המבחן הסטטיסטי:

מבחן WALT:

$$V_t = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta_0 P_t + \beta_1 (D_{1i} P_t) + \beta_2 (D_{2i} P_t) + \beta_3 (D_{3i} P_t) + u_t \quad (\text{U})$$

$$V_t = \alpha + \beta \cdot P_t + u_t \quad (\text{R})$$

אם דוחים את H_0 , יש לבדוק במבחני WALD האם ההבדל הוא בין החותכים או בין השיפועים:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$$

באם דוחים את H_0 יש להמשיך לבדוק באמצעות מבחני t:

$$H_0: \beta_j = 0 \quad H_0: \alpha_j = 0$$

? ידוע כי התוצר בישראל תלוי בתוצר בארה"ב, על כן הוצע לבחון את המודל:

$$gdp_i = \alpha + \beta \cdot gdp_usa_i + u_i$$

כאשר gdp הוא אחוז הצמיחה בתוצר בישראל.

ו- gdp_usa הוא אחוז הצמיחה בתוצר בארה"ב.

הועלתה הטענה שיש עונתיות לפי רבעונים לשם כך הוגדר משתנה הדמי $q = \text{רבעון}$.

רשמו את השערות האפס עבור הטענות הבאות:

א. הרבעונים משפיעים על אחוז הצמיחה בתוצר בישראל באופן כללי וכתלות בהשפעת התוצר האמריקאי?

ב. השפעת הרבעון השלישי על התוצר בישראל זהה להשפעת הרבעון הרביעי?

ג. השפעת התוצר האמריקאי על התוצר בישראל ברבעון השני זהה לרבעון הראשון?

ד. הרבעונים השלישי והראשון משפיעים באופן זהה על התוצר בישראל באופן ישיר וגם כתלות בתוצר האמריקאי.

משתני דמי עבור שני משתנים איכותיים

נתבונן בדוגמא שבה יש שני משתנים איכותיים המשפיעים על פונקציית השכר -

מגדר (אישה, גבר) וגזע (לבן, שחור).

נגדיר משתנה דמי G שיקבל 1 אם מדובר בגבר ו-0 אחרת (אישה).

נגדיר משתנה דמי R שיקבל 1 אם מדובר בלבן ו-0 אחרת (שחור).

נבדוק כיצד מגדר וגזע משפיעים על השכר ההתחלתי (החותך), כאשר השכר תלוי גם בשנות לימוד (S_t).

(1) הבדל בחותך ללא אינטראקציה

המודל:

$$W_t = \alpha_0 + \alpha_1 G + \alpha_2 R + \beta \cdot S_t + u_t$$

במודל זה - אין השפעה משולבת של מגדר וגזע על השכר ההתחלתי.

במילים אחרות, ההבדל בשכר ההתחלתי בין גברים ונשים לא תלוי בגזע (זהה עבור שחורים ועבור לבנים) ולהיפך - ההבדל בשכר ההתחלתי בין לבנים לשחורים לא תלוי במגדר (זהה עבור נשים וגברים).

ניתן לבדוק השערות על כל אחד מהמשתנים הב"ת האיכותיים בנפרד:

1. הבדל בשכר ההתחלתי בין גברים לנשים: $H_0: \alpha_1 = 0$

2. הבדל בשכר ההתחלתי בין שחורים ללבנים: $H_0: \alpha_2 = 0$

(2) הבדל בחותך עם אינטראקציה

המודל: $W_t = \alpha_0 + \alpha_1 G + \alpha_2 R + \alpha_3 G \cdot R + \beta \cdot S_t + u_t$

במודל זה הטענה היא כי קיימת, בנוסף להשפעה של מגדר וגזע בנפרד על השכר, גם השפעה משולבת (אינטראקציה) של מגדר וגזע על השכר ההתחלתי.

במילים אחרות, ההבדל בשכר ההתחלתי בין גברים ונשים תלוי בגזע (שונה אם מדובר בשחורים או בלבנים) ולהיפך.

במודל זה, לעומת הקודם, נוספת ההשערה לבדיקת השפעת האינטראקציה בין מגדר לגזע על השכר ההתחלתי:

$$H_0: \alpha_3 = 0 \quad .3$$

? חוקרת בדקה השפעות השכלה, מגדר וניסיון על הכנסה מעבודה לפי המשוואה הבאה:

$$\ln(MWAGE) = \alpha_0 + \alpha_1 \cdot S + \alpha_2 \cdot E + \alpha_3 \cdot (S \cdot E) + \beta_0 \cdot EXP + \beta_1 \cdot (EXP \cdot S) + \beta_2 \cdot (EXP \cdot E) + \beta_3 \cdot (EXP \cdot S \cdot E) + U$$

כאשר S : משתנה דמי =1 עבור נשים, 0=גברים

E משתנה דמי : 1= עבור השכלה גבוהה (>12 scl), 0=השכלה נמוכה

בטרם ניגשים לפיתרון השאלה יש להכין את טבלת העזר הבאה:

הפרש	השכלה נמוכה (E=0)	השכלה גבוהה (E=1)	
חותך: $\alpha_2 + \alpha_3$ שיפוע: $\beta_2 + \beta_3$	$\alpha_0 + \alpha_1 + (\beta_0 + \beta_1)EXP_t$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + (\beta_0 + \beta_1 + \beta_2 + \beta_3)EXP_t$	נשים (S=1)
חותך: α_2 שיפוע: β_2	$\alpha_0 + \beta_0 EXP_t$	$\alpha_0 + \alpha_2 + (\beta_0 + \beta_2)EXP_t$	גברים (S=0)
	חותך: α_1 שיפוע: β_1	חותך: $\alpha_1 + \alpha_3$ שיפוע: $\beta_1 + \beta_3$	הפרש

א. רשמו את הפונקציה לחישוב:

1. תחזית לוג השכר עבור גבר בעל השכלה נמוכה ו- 10 שנות ניסיון.
2. תחזית לוג השכר ההתחלתי עבור נשים משכילות.
3. לאחר כמה שנות ניסיון ישתווה השכר של נשים משכילות לזה של גברים משכילים?

ב. רשמו את השערות האפס המתאימות לבדיקת הטענות הבאות:

1. אין השפעה של מגדר והשכלה על השכר.
2. השפעת ההשכלה אינה תלויה במגדר.
3. אין השפעות השכלה אצל גברים.
4. אין הבדל בשיעורי התשואה לניסיון, בקרב הנשים.

שאלות ממבחנים-רגרסיה

מבחן מס' 1

שאלה 1

להלן תוצאות הרצת רגרסיה של Y בתלות ב- X עבור 10 תצפיות (חלק מהנתונים הושמטו בכוונה מהפלט, אך ניתנים לחישוב על ידך).

$$\text{נתון כי: } \sum (X_i - \bar{X})^2 = 1475.6$$

מקור	סכום ריבועים SS
רגרסיה	$SSR = 2148.6$
שאריות	$SSE = ?$
סה"כ	$SST = ?$

משתנה	מקדם	טעות תקן	ערך סטטיסטי(מתוקן)	p-value מובהקות
	b_i	S_{b_i}	t	
קבוע (חותך)	-24.7	11.3	?	?
X	1.20	?	10.5	

א. מהו SST?

א. לא ניתן לקבוע.

ב. 1994.42

ג. 2304.1

ד. 1629.78

ב. האם הרגרסיה מובהקת? בדוק לפי p value

א. הרגרסיה מובהקת

ב. הרגרסיה אינה מובהקת.

שאלה מס' 2

לפניך פלט רגרסיה פשוטה (ממנו הושמטו נתונים שבאפשרותך להשלים), המתאר את ציון המבחן כפונקציה של מספר התרגילים שהגיש הסטודנט במהלך הסמסטר, ידוע כי כל הנחות המודל תקפות.

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.842105598
R Square	0.709141839
Adjusted R Square	0.688366256
Standard Error	5.315523758
Observations	16

ANOVA

	<i>Df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	964.4329004	964.4329004	34.13343	4.28E-05
Residual	14	395.5670996	28.25479283		
Total	15	1360			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	50.99134199	4.318918368	11.80650747	1.15E-08
מספר התרגילים	4.086580087	0.699471573	5.842381939	4.28E-05

1. ע"פ הנתונים, אחוז השונות של ציוני המבחן המוסברת ע"י מספר התרגילים שהגיש הסטודנט, היא _____ . אם נוסף משתנים נוספים, אחוז השונות המוסברת _____, ו- $R_{adjusted}$ _____ .

א. 84% יגדל, לא ניתן לקבוע ללא נתונים נוספים.

ב. 84% , יקטן, יגדל.

ג. 70.9% , יגדל, לא ניתן לקבוע ללא נתונים נוספים.

ד. 70.9% , יגדל, יקטן.

ה. 68.8% יגדל, יגדל.

ו. 68.8% יקטן, יקטן.

ז.

2. מהו הרב"ס של שיפוע הרגרסיה β_1 ? (בר"מ של 1%)

א. $2.58 < \beta < 5.58$

ב. $2 < \beta < 6.17$

ג. $1.74 < \beta < 6.88$

ד. $2.86 < \beta < 5.3$

במטרה לנבא בצורה טובה יותר את הצלחת הסטודנטים בבחינה, החליט החוקר להוסיף 2 משתנים נוספים לניתוח הרגרסיה.

מספר השיעורים בהם נכח הסטודנט, ומספר השעות שלמד לבחינה.

לפניכם הפלט החסר:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.880163577
R Square	0.774687922
Adjusted R Square	0.718359902
Standard Error	5.053253294
Observations	16

ANOVA

	df	SS	MS	F
Regression	3	?	351.1918579	13.75315
Residual	12	306.4244263	25.53536886	

Total 15 ?

	Coefficients	Standard Error	t Stat	P-value
Intercept	37.08959571	8.726136945	4.250402663	0.001127
מספר השיעורים	2.610000755	1.429904588	1.82529714	0.092932
מספר התרגילים	3.198068014	1.239162836	2.58082951	0.024061
שעות לימוד לבחינה	-0.108373802	1.021202927	-0.106123669	0.917238

1. בדוק את ההשערה כי הרגרסיה מובהקת לכל אחד מהמשתנים המסבירים בר"מ של 1%.

א. הרגרסיה אינה מובהקת לכל המשתנים שנבדקו.

ב. לא ניתן לקבוע מהנתונים האם הרגרסיה מובהקת.

ג. הרגרסיה מובהקת למשתנה מספר התרגילים, אך אינה מובהקת למשתנים מספר השיעורים ומספר שעות הלימוד לבחינה.

ד. הרגרסיה מובהקת למשתנה מספר התרגילים ומספר השיעורים, אך אינה מובהקת למשתנה שעות הלימוד לבחינה.

2. מהו SSR של הרגרסיה המרובה?

א. $SSR=964.4$

ב. $SSR=1053.57$

ג. $SSR=694.57$

ד. $SSR=853.57$

ה. לא ניתן לחשב את SSR מהנתונים שהתקבלו.

מבחן מס' 2

שאלה מס' 1

הועלתה השערה שהוצאות האחזקה של מערכת לעיבוד נתונים קשורות למספר שעות השימוש השבועיות במערכת. להלן תוצאות חלקיות של פלט EXCEL של ניתוח רגרסיה בין- Y הוצאות אחזקה שנתיות (במאות \$) ו- X מספר שעות השימוש השבועיות.

SUMMARY OUTPUT

<i>Regression Statistics</i>					
Multiple R					
R Square					
Adjusted R Square					
Standard Error					
Observations 10					

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression		860.051			0.00012
Residual					
Total		1004.525			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	10.528	3.745		0.023
שעות שימוש	0.953		6.901	

א. לאור התוצאות ה- p-value בבדיקת ההשערה:

$$H_0: \beta_1 \leq 0.5$$

$$H_1: \beta_1 > 0.5$$

1. $Pv < 0.005$

2. $0.005 < pv < 0.01$

3. $0.01 < pv < 0.025$

4. $0.025 < pv < 0.05$

5. $Pv > 0.05$

ב. אם היינו מריצים רגרסיה בה- Y מספר שעות השימוש השבועיות ואילו המשתנה המסביר- X, הוצאות אחזקה שנתיות (במאות \$), אזי השיפוע של קו הרגרסיה יהיה:

1. 0.953

2. 1.049

3. 10.528

4. 0.095

5. 0.898

שאלה מס' 2

הועלתה השערה שמספר התקלות ברכב Y קשורה לגיל הנהג X. לשם כך נלקח מדגם של 10 נהגים.

כמו כן חושבו הסכומים הבאים:

$$\sum X_i^2 = 14,227 \quad \sum X_i = 363 \quad \sum Y_i = 13 \quad \sum Y_i^2 = 29 \quad \sum X_i Y_i = 366$$

משוואת קו הרגרסיה נתונה על ידי: $\hat{Y} = 4.96 - 0.1X_i$

א. ערכו של מקדם המתאם הלינארי בין מספר התקלות לבין גיל הנהג הוא:

1. -10.59

2. -0.9395

3. 0.8826

4. -0.1

החוקר לא היה מרוצה מעוצמת הקשר ולכן החליט להוסיף לרגרסיה את המשתנים הבאים :

מספר הק"מ שהמכונית נסעה (באלפי ק"מ) וסוג הרכב. במדגם נכללו 2 סוגי רכבים: A ו B

כאשר סוג רכב B קודד כערך 0 וסוג רכב A קודד בערך 1.

להלן פלט הרגרסיה המרובה. שימו לב כי חלק מהערכים בפלט חסרים:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	
R Square	
Adjusted R Square	
Standard Error	0.204047
Observations	10

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Regression	3			0.00002
Residual	6	0.249811		
Total	9			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	1.407943	0.71094		
X1 גיל הנהג	-0.03094	0.014611		
X2 סוג הרכב	0.239574	0.152994		
X3 ק"מ באלפים	0.027666	0.005329		

ב. להלן מספר טענות לגבי מובהקות הרגרסיה ומובהקות המשתנה סוג הרכב ברמת מובהקות 0.1:

1. הרגרסיה מובהקת והמשתנה סוג הרכב מובהק ברמת מובהקות 0.1.
2. הרגרסיה מובהקת, אך המשתנה סוג הרכב אינו מובהק ברמת מובהקות 0.1.
3. הרגרסיה אינה מובהקת, אך המשתנה סוג הרכב מובהק ברמת מובהקות 0.1.
4. הרגרסיה והמשתנה סוג הרכב אינם מובהקים ברמת מובהקות 0.1.

ג. SST בפלט הרגרסיה המרובה שווה ל-

12.1 1.

16.0 2.

20.0 3.

4. אין מספיק נתונים לחשבו.

ד. לאור התוצאות, רב"ס למקדם המשתנה מספר הקילומטרים, בר"מ 5%:

1. (0.015,0.04)

2. (0.017,0.038)

3. (0.02,0.035)

4. אין מספיק נתונים לחשבו.

ה. אם היינו מקודדים את סוג רכב B כערך 1 וסוג רכב A קודד בערך 0 משואת הרגרסיה הייתה:

1. נשאר ללא שינוי

2. $\hat{Y} = 1.4079 - 0.0309X_{1i} - 0.2395X_{2i} + 0.0276X_{3i}$

3. $\hat{Y} = 1.6475 - 0.0309X_{1i} - 0.2395X_{2i} + 0.0276X_{3i}$

4. לא ניתן לדעת ללא הרצה מחדש

ו. החוקר רצה להוסיף משתנה מסביר נוסף- X4 מספר השנים שחלפו מאז קבלת רישיון הנהיגה. להלן פלט הרגרסיה (חלק מהנתונים חסרים) עם 4 המשתנים המסבירים:

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.5	0.71		
X1 גיל הנהג	5.1	14.1		
X2 סוג הרכב	0.25	0.2		
X3 ק"מ באלפים	0.02	0.008		
מס' השנים שחלפו מאז קבלת רשיון X4	-5.13094	0.8		

על סמך נתונים אלו:

1. חשש סביר למולטיקוליניאריות
2. התחזית הנקודתית של מספר התקלות תהיה דומה לזו של הרגרסיה הקודמת (עם 3 המשתנים המסבירים).
3. קיימת קורלציה גבוהה בין חלק מהמשתנים המסבירים.
4. כל התשובות נכונות.

שאלה מס' 3

במפעל מסוים הורץ מודל של רגרסיה ליניארית פשוטה על 8 עובדים כאשר Y - תפוקת העובד ו- X - גיל העובד. נמצא שהחלק המוסבר על ידי הרגרסיה הוא 74.

הטעויות שהתקבלו מופיעות בחלקן בטבלה שלהלן:

e_7	e_6	e_6	e_5	e_4	e_3	e_2	e_1
0	2	-2	2	?	2	-3	0

אחת מהטענות שלהלן נכונה:

1. מקדם ההסבר בין X ל-Y הוא 0.74

2. לא ניתן לחשב את מקדם המתאם המרובה

3. $SSR=18$

4. $SSE=74$

5. אף אחת מהטענות איננה נכונה

שאלה מס' 4

חברה בדקה את הקשר בין המכירות השבועיות לבין הוצאות הפרסום. מנהל השיווק בדק מדגם של 30 שבועות ומצא כי קו הרגרסיה לניבוי המכירות השבועיות על סמך הוצאות הפרסום הוא:

$$\hat{y} = 1.25x - 0.5$$

איזה מהמשפטים הבאים אינו אפשרי?

- א. קיים קשר מושלם בין המכירות השבועיות לבין הוצאות הפרסום.
- ב. קיים קשר חזק וחיובי בין המכירות השבועיות לבין הוצאות הפרסום.
- ג. קו הרגרסיה לניבוי הוצאות הפרסום על סמך המכירות השבועיות מתלכד עם הקו הנתון.
- ד. קו הרגרסיה לניבוי הוצאות הפרסום על סמך המכירות השבועיות אינו מתלכד עם הקו הנתון.